Int J Appl Earth Obs Geoinformation 64 (2018) 354–364

Contents lists available at ScienceDirect

Int J Appl Earth Obs Geoinformation

journal homepage: www.elsevier.com/locate/jag

Comparing experts and novices in Martian surface feature change detection and identification

Jessica Wardlaw[a,*], James Sprinks[a], Robert Houghton[a], Jan-Peter Muller[b], Panagiotis Sidiropoulos[b], Steven Bamford[a], Stuart Marsh[a]

[a] University of Nottingham, Nottingham, UK
[b] Imaging Group, The Mullard Space Science Laboratory, University College London, UK

ARTICLE INFO

Keywords:
Crowd sourcing
Citizen science
Volunteered geographic information
Planetary science
Change detection
Image analysis

ABSTRACT

Change detection in satellite images is a key concern of the Earth Observation field for environmental and climate change monitoring. Satellite images also provide important clues to both the past and present surface conditions of other planets, which cannot be validated on the ground. With the volume of satellite imagery continuing to grow, the inadequacy of computerised solutions to manage and process imagery to the required professional standard is of critical concern. Whilst studies find the crowd sourcing approach suitable for the counting of impact craters in single images, images of higher resolution contain a much wider range of features, and the performance of novices in identifying more complex features and detecting change, remains unknown.

This paper presents a first step towards understanding whether novices can identify and annotate changes in different geomorphological features. A website was developed to enable visitors to flick between two images of the same location on Mars taken at different times and classify 1) if a surface feature changed and if so, 2) what feature had changed from a pre-defined list of six. Planetary scientists provided "expert" data against which classifications made by novices could be compared when the project subsequently went public.

Whilst no significant difference was found in images identified with surface changes by expert and novices, results exhibited differences in consensus within and between experts and novices when asked to classify the type of change. Experts demonstrated higher levels of agreement in classification of changes as dust devil tracks, slope streaks and impact craters than other features, whilst the consensus of novices was consistent across feature types; furthermore, the level of consensus amongst regardless of feature type. These trends are secondary to the low levels of consensus found, regardless of feature type or classifier expertise. These findings demand the attention of researchers who want to use crowd-sourcing for similar scientific purposes, particularly for the supervised training of computer algorithms, and inform the scope and design of future projects.

1. Introduction

Detection of change in satellite images of Earth and other planetary bodies is of significant scientific interest in the monitoring of environmental and climate change. Automating the detection of surface features over different spatial and temporal scales, however, remains complex and computationally expensive. Variation in the quality and coverage of images render them difficult for computers to process, in addition to the atmospheric and morphological influences on the "visibility" of features (Kim et al., 2005). Although we anticipate the development of increasingly subtle and powerful image processing and machine learning systems (Sidiropoulos and Muller, 2016), there remains a role for the human analyst particularly when variability is emphasised and human aptitudes of flexibility and judgement are called into play (e.g. interpreting rare events or features to make serendipitous discoveries). However, there is currently a clear, growing and profound imbalance between the number of expert observers and the sheer volume of satellite data available to the wider scientific community (See et al., 2016). One solution to this is to crowdsource analysis of imagery – a process often discussed within the realm of Citizen Science (Bonney et al., 2009). However, the viability of this solution rests on the fundamental question of whether a collection of suitably equipped amateurs can generate data of comparable quality to that produced by experts (Salk et al., 2016).

This paper investigates the potential power of novices to address two challenges that face the future application of a crowd-sourcing approach for the analysis of satellite imagery: detection of a wider range of surface features and changes in the appearance of these

* Corresponding author.

http://dx.doi.org/10.1016/j.jag.2017.05.014
Received 31 October 2016; Received in revised form 28 March 2017; Accepted 29 May 2017
Available online 12 June 2017

features that reflect dynamic changes on the surface. Crowdsourcing has successfully classified surface features in Earth Observation, through calibration with ground truth (Zhao et al., 2014; See et al., 2016). The crowd is commonly used to count craters for estimating the age of lunar surfaces, a task which implicitly assumes that craters can be reliably identified, and further relies on measurements of crater diameter for age calculations (Robbins et al., 2014). In lunar images factors such as atmospheric distortion and the range of surface features are reduced so that the effects of human subjectivity can be isolated (Gault, 1970; Kirchoff et al., 2011). Robbins et al. (2014) investigated the consistency of expert classifications of craters in relation to terrain type, size and frequency, across different user interfaces. For all variables, only annotations of the smallest craters (< 10 pixels in diameter) were significantly different. They concluded "volunteers are approximately as good as experts in identifying craters…so long as enough volunteers examine the image to derive a robust result," with the caveat that accuracy for any single crater or cluster of craters is not important (Robbins et al., 2014; 126). Comparison of automated feature detection with the subjectivity introduced by humans has found differences between and within the classifications of individuals, for example on different days (Tar and Thacker, 2016). Successful cataloguing of geological landmarks could facilitate the filtering of imagery according to features of interest but the future utility of any automated process for this would require a significant human effort to label examples for training the algorithm (Wagstaff et al., 2012; Wagstaff et al., 2015). Whilst the work of Robbins suggests that novices can produce comparable annotations of impact craters to experts, their ability to identify other surface features of interest remains untested.

The present study extends previous work to detecting changes in images of the surface of Mars, in which features change at different rates, from rapidly moving dust devils, seasonal and inter-annual fluctuations of the polar ice caps and recurring slope lineae (indicating contemporary water activity), and slowly shifting sand dunes. Scientific interest in detecting changes in features such as impact craters (Kim et al., 2005; Bue and Stepinski, 2007; Li et al., 2015), gullies (Stepinski and Collier, 2004) and sand dunes (Bandeira et al., 2013) on Mars is high because changes reveal the evolution of the climate and geology of the planet; repeat image coverage for change detection is increasingly available and, until surface data can be validated with any certainty, alternative approaches are needed.

Although beyond the scope of this study, the introduction of human analysts, even within the context of the crowd-sourcing approach, brings into play other potential confounds on performance. Visual search is known to be affected by feature complexity (Lloyd and Hodgson, 2002) and size (Warner et al., 2015), scene context (Castelhano and Heaven, 2010), information density and presentation method (Chang et al., 2012), in addition to the human factors associated with performing visual search for a prolonged period of time (See, 2012). Change detection studies are also relevant in this context (Rensink, 2002), as well as those concerned with the quality of Volunteered Geographic Information (Haklay, 2010; Foody et al., 2013).

The ultimate goal of the on-going development of the algorithm is to achieve fully automated change detection and characterisation. Such a task is typically tackled with a supervised learning approach using a ground-truth dataset, but no publicly available ground-truth currently exists for planetary surfaces. The crowdsourcing this paper presents is thus intended to produce annotations for developing a fully automatic change detection algorithm. More information about the co-registration and the change detection algorithm can be found in Sidiropoulos and Muller (2016).

Section 2 now sets out the approach used to study these questions. Section 3 will present the consensus found within and between novices' and experts' classifications of change, and feature type that changed. Section 4 will discuss key findings and their implications for the remote sensing community, and designers of crowd-sourcing platforms for the classification of geomorphological features.

2. Method

2.1. Experimental design

To investigate novice performance in detecting 1) more complex features and 2) changes in features over time, this work presents the results of a Citizen Science project built with the project builder 'Panoptes' on Zooniverse.org and tested with experts and novices to directly their classifications of dynamic geological changes in Martian images, with a task designed for participants to compare two images of the same location but at different times (Bowyer et al., 2015).

The current interest in Martian exploration and the volume of images that have amassed since the planet was first imaged forty years ago represent an outstanding opportunity for the investigation presented. The images under study were processed from genuine images of the surface of Mars, so that participants would not anticipate what they would see. Prior to public release, doctoral Planetary Science students and post-docs classified images within a workshop at University College London's (UCL) Mullard Space Science Laboratory. Their exclusive access over the two days enabled separation of their "expert" classifications from those of volunteer "novices" over the following months.

2.2. Apparatus/materials

The study used images extracted from high-resolution image strips acquired by four orbital cameras described in Table 1.

First, the raw images were projected, or "co-registered", to a single coordinate system, to enable comparison. Since no high-resolution global datum exists for Mars, a mix of High-Resolution Stereo Camera (HRSC) Orthorectified Images (ORI) and Digital Terrain Models (DTMs), covering almost 50% of Mars, was selected for use as a baseline (Sidiropoulos and Muller, 2015). The co-registration technique was developed to achieve a fast and fully automatic co-registration of large volumes of data for generating an abundant input for change detection (Sidiropoulos and Muller, 2016). The subsequent set of co-registered images comprised of overlapping image pairs, which were then processed by an algorithm for detection of "regions-of-interest" (Sidiropoulos and Muller, 2016). The algorithm selected 868 regions-of-interest, each 512 × 512 pixels in size, as surface change candidates.

The change detection algorithm used is a "late fusion classification scheme" (Ye et al., 2012), and defines four types, or "classifiers", of change. Each classifier models a distinct type of surface change and produces a single, independent output in the form of a "confidence score" (Ye et al., 2012) from 0 to 1 for the probability of a positive classification, with 1 meaning 100% certainty that a pair of images

Table 1
Description of the cameras that took the images used in this study.

Camera	Spacecraft	Dates of Operation	Resolution	Reference
Context Camera (CTX)	Mars Reconnaissance Orbiter	2006-present	6m/pixel	Bell et al. (2013)
High-Resolution Stereo Camera (HRSC)	Mars Express	2004-present	12.5m/pixel	Jaumann et al. (2007)
Thermal Emission Imaging System (THEMIS)	Mars Odyssey	2002-present	17.5m/pixel	Christensen et al. (2004)
Mars Orbiter Camera − Narrow Angle (MOC−NA)	Mars Global Surveyor	1997–2006	1.5-12m/pixel	Malin et al. (2010)

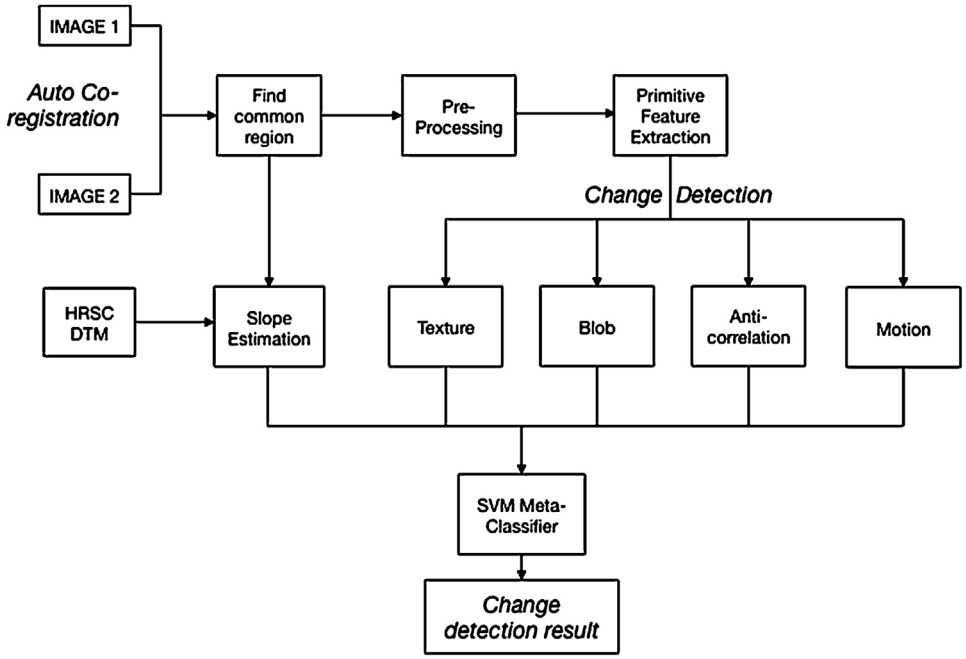

Fig. 1. The flowchart of the change detection algorithm. Details of the auto co-registration of image pairs and creation of the High Resolution Stereo Camera Digital Terrain Model (HRSC DTM) are documented in Sidiropoulos and Muller (2016) and Gwinner et al. (2016) respectively.

includes a change of this type. The results of these classifiers are combined by means of a secondary "meta-classifier", which generates a final score that is compared to a threshold to determine the presence of change. The flowchart of this scheme can be found in Fig. 1.

Classifiers were defined according to visual characteristics and not according to scientific context, i.e. classifiers were not defined to directly map to features on the Martian landscape (Fig. 1). Classifiers were designed to find four types of change:

- *Texture* changes, which are identified in images in which the surface texture has changed (e.g. created by Aeolian activity).
- *Binary large object (Blob)* changes, attributed to image pairs in which an approximate homogenous patch only appears in one of the two images and are used here as proxies for new large-sized features (e.g. slope streaks).
- *Anticorrelation* changes, which are identified by negative spikes in the image pair correlation and are used here as proxies for new small-sized features that emerge (e.g. a new impact crater).
- *Motion*-type changes, which are found through the detection of a local mis-registration between the two images.

To increase certainty, a final classifier estimates and compares the surface slope in images since they are co-registered and ortho-rectified to the High Resolution Stereo Camera (HRSC) Digital Terrain Model (DTM).

This classification scheme is based on supervised learning, which requires training with both positive and negative annotations. Such annotations are currently sparse, since large-scale studies of changes in geomorphological features on Mars are unavailable. Therefore, the approach we use includes a feedback loop between the automatic change detection results and the crowdsourcing annotations. More specifically, this study used a preliminary automatic change detection algorithm, which used a small set of manually annotated images to estimate classification parameters. Subsequently, the crowdsourcing annotations are used to train the automatic change detection scheme, which then produce a second round of lower false positive rate, while possibly more repeats will happen in the future.

A project was built on Zooniverse.org to run the test. Zooniverse is a consortium of researchers from the Adler Planetarium in Chicago, USA and the Department of Physics at the University of Oxford. It uses Amazon Web Services to host a plethora of Citizen Science projects, spanning fields from Astronomy to Zoology. A project builder interface negates the need for coding and frees researchers to create their own projects for volunteers to classify or analyse images according to their needs.

Crucially, the project builder manages the order in which volunteers classify images. In general it applies rules so that no individual volunteer sees the same image pair more than once if they are registered and logged in; if they are not logged in it will randomly select an unretired image from across the images that remain.

2.3. Participants

22 Planetary Scientists attended a workshop on 3D data and were invited to participate in the experiment; the group mostly comprised PhD students and post-doctoral researchers funded by the Europlanet 2020 Research Infrastructure, a European Commission Horizon 2020 project to integrate and support Planetary Science activities across Europe. The requirements for participation in the workshop ensured that participants had the necessary planetary imagery expertise to provide "expert" ("gold standard") data in place of ground truth; this paper will now refer to these participants as "experts" for clarity.

After the workshop, the project was launched informally on social media and local email networks to collect data from volunteers. When five different volunteers had independently classified an image pair, the image pair was removed, or "retired" from any further analysis in a tradeoff between having enough data to compare with the Planetary Scientists' annotations for any one pair of images, and the need to analyse as much of the planet's surface as possible. Data collection for the present study stopped after four months, when enough data had been volunteered.

2.4. Procedure

Experts signed an information sheet and consent form to ensure that they understood the task, why they were doing it, and gave their permission to use their classifications for stated purposes; furthermore, they could leave the study and/or request removal of their data at any time. The experiment was described and participants registered on www.zooniverse.org during a half an hour session before a lunch break. This had two benefits: it gave attendees time to consider their participation and ask questions about it during the lunch break, but also

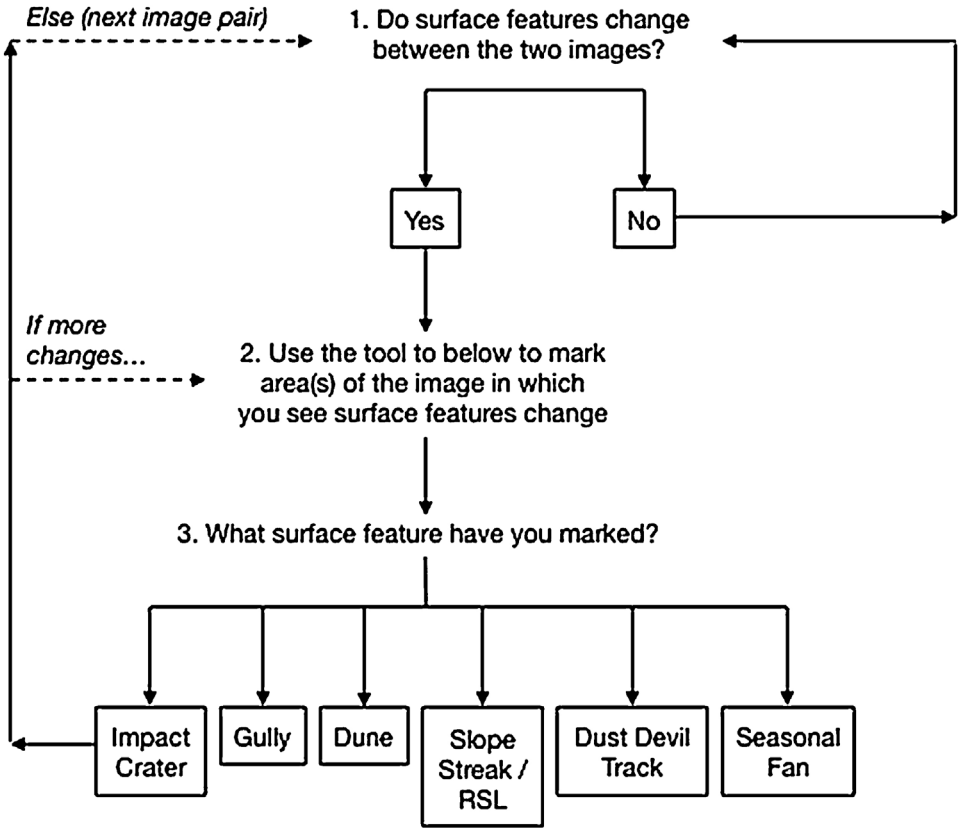

Fig. 2. Task workflow: Mars in Motion.

ensure participants were ready to start the task together.

The workflow depicted in Fig. 2 shows that participants began the task by inspecting two images of the same location on the surface of Mars at different times and selecting whether or not features had changed (Fig. 3).

If they were unsure what type of changes to annotate they could click on a *'Need some help with this task?'* button for a hint (Fig. 4). The help information was deliberately designed to provide only a high-level hint so that participants would understand what changes they should mark but also use their judgement.

If participants marked a change, a subtask asked them to select which type of surface feature they had just marked, from a list of six: impact crater, gully, dune, slope streak or recurring slope lineae, dust devil track and seasonal fan (Fig. 5). The presentation of multiple features types at this stage of the task was a deliberate design decision to make the task less repetitive and to mitigate the detrimental effect of fatigue (See, 2012).

These features were chosen for their scientific interest, and the frequency with which they appear on the surface of Mars. A field guide provided examples of change for each feature type to assist classification (Fig. 6).

Expert participants were encouraged to provide feedback on their experience of the website via a semi-structured online survey after using the website for one hour; they were also invited to take notes anonymously on post-it notes during the experiment so that they could note their thoughts as they occurred. The purpose of this was to provide context for the data during analysis. In contrast, public participants were not asked to complete the feedback survey or restricted to one

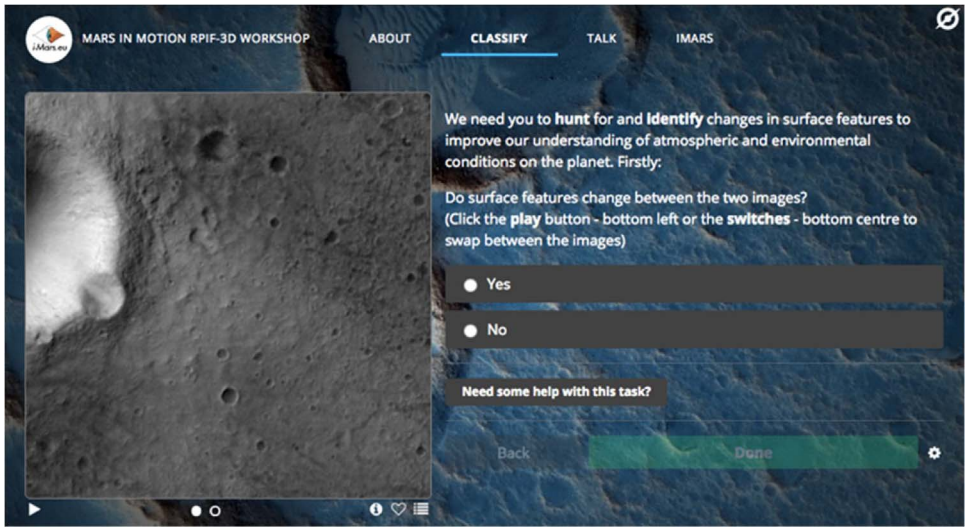

Fig. 3. The task for workshop participants (Step 1 in Fig. 2).

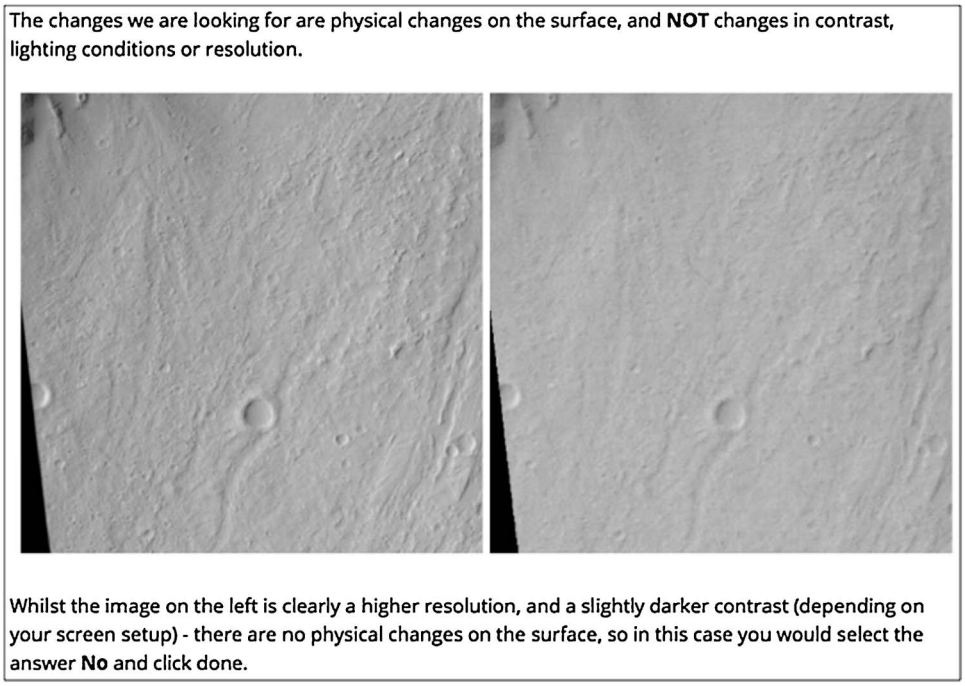

Fig. 4. The 'help' text provided with the first part of the task.

hour of participation; their participation was discretionary and not under the controlled conditions of the experts, so they could return to the website and classify as many image pairs as they liked without seeing the same image pair twice.

2.5. Data captured

The Zooniverse.org collects classification data automatically each time the "Done" button is clicked and can be downloaded by the creators of the project at any time.

For the purposes of this study, three sets of data were of interest: whether or not a change was seen, the type of feature(s) marked as changed, and the time taken to make the classification. Analysis did not directly compare the location of feature annotations on the image, but instead considered 1) the proportion of people who saw a change between images, and 2) the type of surface feature participants labelled. Importantly, analysis was restricted to images seen by more than one person in order to calculate a consensus for an image pair.

Classification data also includes the start and finish time for each individual classification to millisecond accuracy. These were used to calculate task time and explored to determine whether or not the expert and novice classifiers spent a similar length of time classifying each image pair and extend comparison of their performance.

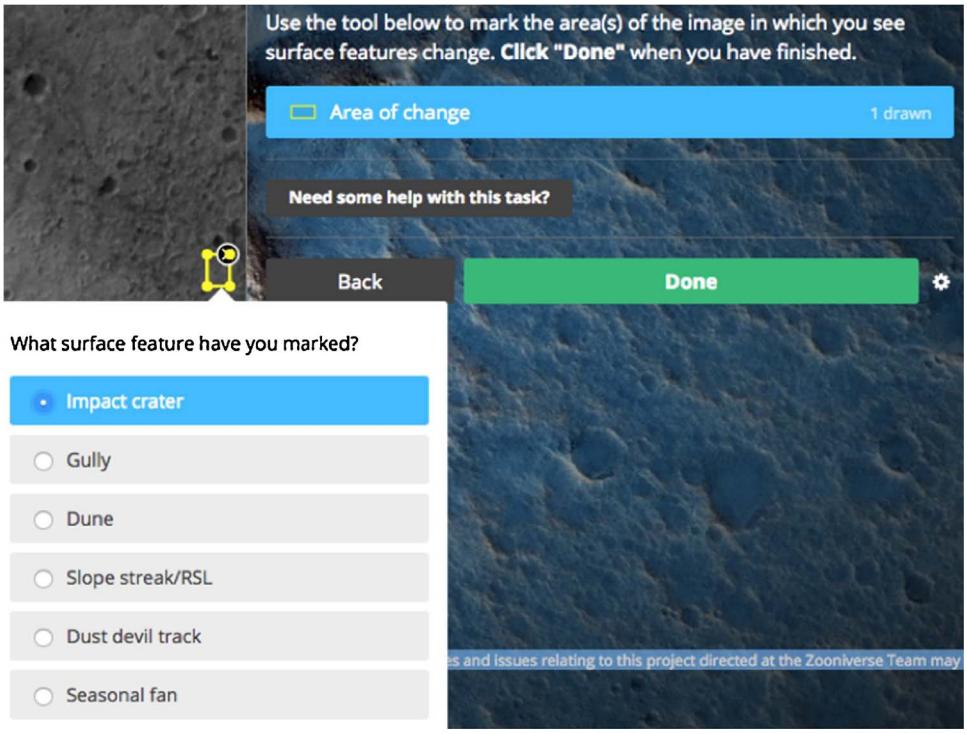

Fig. 5. The subtask of feature identification (Step 3 in Fig. 1).

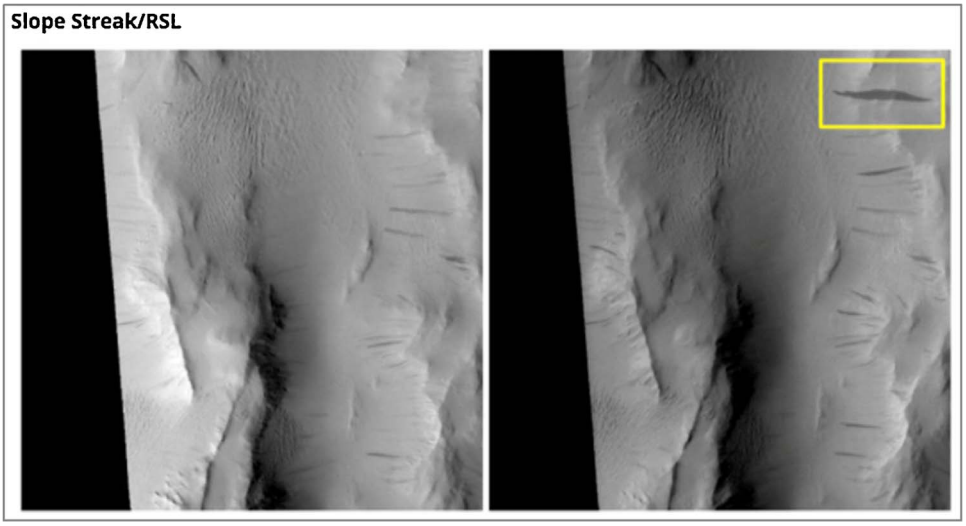

Fig. 6. Example to demonstrate the type of feature change that should be marked. If participants did not feel that any of the listed features matched what they had marked then they could amend their classifications, until they clicked "Done", which logged their classifications and presented participants with a new pair of images.

2.6. Consensus analysis

This data was collected with the objective of calculating a measure of classification consensus, which the crowdsourcing approach uses to assess the accuracy and trustworthiness of volunteered data; the higher the level of agreement, the more confidence with which researchers can use its analysis.

In the study presented classification consensus was first calculated to determine if levels of agreement on seeing a change (i.e. Step 1, Fig. 3) varied between the two groups. Fig. 7 shows this calculation, which was carried out for each image pair for novice and expert classifications independently, where two or more people from one of those groups had classified an image (Fig. 8).

Consensus analysis was then extended to calculate the consensus for change with each type of landscape feature (e.g. crater, gully, dune) for every image pair seen by more than one person. For each type of landscape feature, consensus was defined as the percentage of people who marked one or more changes on the image pair with the total number of views of the image pair (Fig. 7). For example, consider an image pair seen by five people. Three people marked changes in slope streaks. Regardless of the number of slope streaks they marked, this image pair would have a consensus of 60% for slope streaks.

With a consensus calculated for each type of change (e.g. crater, gully, dune) for each image pair, the mean consensus for each type of change was calculated across all images in which each type of change had been marked. The mean consensus for each type of change could then be compared to assess the relative difficulty of spotting a change in each type of landscape feature.

3. Results

This section reports the analysis of classifications provided by both experts and novices in three parts. First we consider the experts' classifications and summarise them with details of how many image pairs they analysed, how many of these had changed and calculate the consensus within their classifications for changes in each feature type. Second, the classifications provided by novices were analysed in the same way. Finally, this section compares agreement between novices and experts' classifications of change for each feature type.

3.1. Consensus regarding 'is there change?': experts vs novice

First, the classification consensus amongst expert and novices is compared for the first part of the task, presented in Fig. 3, for each image pair. In total, two or more people from within each group classified 1301 image pairs; two or more novices classified 738 image pairs and two or more experts classified 563 image pairs. A Mann-Whitney *U* test on the consensus for change on these 1301 image pairs revealed no significant difference in the level of agreement for change between expert and novice classifiers ($U = 197605$, $p = 0.078$).

3.2. Consensus of experts

In total, experts contributed 1877 classifications during the time allotted in the two-day workshop, of which 1553 (82.7%) were classifications of 'no change' and 324 (17.3%) were classifications of surface changes. Of the full set of 868 image pairs, experts classified 783 unique images pairs, of which 580 (25.9%) had no changes annotated. In order to calculate consensus, 220 image pairs that were only seen once were eliminated from analysis. Of the remaining 563 unique image pairs, seen by more than one expert, 175 were marked with change, with the mean% consensus for change 55.9% (standard deviation = 8.3, standard error = 2.1).

Table 2 describes the data associated with these image pairs according to feature type. The first two rows show the percentage of the whole image set that experts marked with each type of change, followed by the number of image pairs marked with each type of change. The mean consensus is then presented for each type of change amongst experts.

$$\text{Consensus} = \text{nChange}/\text{NN} \times 100$$

where:

N = Total number of people who classified an image pair

nChange = Number of people who annotated an image pair with a change

Fig. 7. Consensus Calculation.

Table 2
Consensus amongst experts with a change detection task on 563 image pairs seen by more than one.

Feature	Crater	Gully	Dune	Slope Streak	Dust Devil	Seasonal Fans
% Image pairs with no change	95.20	98.22	94.14	91.83	88.28	96.80
% Image pairs with change	4.80	1.78	5.86	8.17	11.72	3.20
No. of image pairs with change	27	10	33	46	66	18
Mean% consensus	**38.55**	**32.50**	**35.35**	**55.92**	**63.81**	**37.41**
Standard Deviation	16.61	9.98	12.36	26.17	26.80	12.26
Standard Error	**3.20**	**3.15**	**2.15**	**3.86**	**3.30**	**2.89**

3.3. Novice consensus

In total, volunteers contributed 2919 classifications in the first four months of the project's release, of which 2446 (83.8%) were classifications of 'no change'; the remaining 473 (16.2%) were annotated with surface changes. For image pairs seen by more than one person there were 2834 individual classifications, of which 2385 (84.2%) were entries of 'no change' and 449 (15.8%) were annotations of change.

Novices classified 823 unique images pairs from the full set of 868, of which 531 (64.5%) were not annotated with changes. However, for calculating consensus, analysis excluded 85 image pairs that were only seen by one person. Of 738 unique image pairs seen by more than one person, 445 (60.3%) were identified as having 'no changes'; for the remainder (identified by at least one person as an area in which the surface changed), the mean consensus for change was 41.1% (± standard deviation 22.1, standard error 1.3). Table 3 presents statistics for these image pairs and Fig. 8 illustrates the consensus reached by volunteers with example image pairs..

3.4. Comparing the consensus of experts of novices

These data can be used to compare how the consensus of experts and novices compares for different features types. The means and standard errors from Table 2 and Table 3 are plotted in Fig. 9, which graphically illustrates their differences.

Fig. 9 shows that the most significant differences between the two groups of classifiers are found in annotations of slope streaks and dust devil tracks, both of which are linear in their morphology; the third most significant difference is found in the marking of changes in craters, which goes against the results of Robbins et al. (2014). There are two possible conclusions regarding these features that have the least agreement; either 1) more eyes are better, or 2) the smaller the agreement, the harder and more subjective changes in the feature are for untrained users to detect.

3.5. Agreement between experts and novices

The next stage of analysis directly compares expert and novice classifications of the same image pair. Of the 783 unique image pairs viewed by more than one expert, a subset of 434 image pairs were also viewed by more than one novice in the four months that followed. This subset of images is now used to compare the performance of volunteers ("novices") with Planetary Scientists ("experts"). Table 4 describes, for each feature type, how many of the image pairs experts marked with a change, and the proportion of those that novices marked with the same change, to directly compare the consensus of experts and novices for the same individual image pairs. The first two columns of Table 4 show the overall rate at which the two distinct groups marked changes in each type of surface feature; the final two relay the proportion of images marked by experts with a change, which were marked with the same type of change by novices.

3.6. Task time: experts vs novices

Raw task time data was copied into SPSS from Microsoft Excel for a more detailed statistical analysis of the time spent on the task by experts and novices for the 434 image pairs that had been seen by more than one expert and more than one novice.

The mean and standard deviation of the task time for these 434 image pairs were tested for a normal distribution, for expert and novices independently, to ascertain whether they should be compared with a parametric or non-parametric test and results indicated that neither group was normally distributed (Table 5).

On this basis, the non-parametric Wilcoxon Signed Ranks test was carried out for differences in task time between novices and experts for the same image pair. The results ($Z = -0.519$ based on negative ranks, and asymptotic significance 2-tailed p-value = 0.603) indicated no significant difference.

4. Discussion

Analysis of the results presented in the previous section, and their implications, cannot be discussed before factors that must be considered are described.

Second, workshop participants' feedback, via the online survey, expressed a particular concern that the quality of many images was not sufficient to be able to discern that a feature had changed, and the interface did not allow them to report the poor quality of image pairs. If participants were unsure whether to mark a change or not, their comments suggested that they erred on the side of caution and tended to answer that there was no change in the imagery. Such uncertainty could arise due to artefacts in the images or spotting a feature that did not match the categories offered and could bias classifications towards no changes. Live Citizen Science projects, however, cannot practically control for image quality, since it is subjective and in the case presented was controlled by keeping the resolution of images constant. Future projects might trial a button for participants to click when the image is of poor quality, but the use of such a button must be judicious in case

Table 3
Consensus amongst novices with a change detection task on 738 image pairs seen by more than one.

Feature	Crater	Gully	Dune	Slope Streak	Dust Devil	Seasonal Fans
% Image pairs with no change	92.41	92.68	90.38	89.70	88.48	96.21
% Image pairs with change	7.59	7.32	9.62	10.30	11.52	3.79
No. of image pairs with change	56	54	71	76	85	28
Mean% consensus	**29.31**	**31.07**	**30.13**	**32.98**	**38.00**	**32.24**
Standard Deviation	12.90	10.70	17.08	14.64	21.74	12.00
Standard Error	**1.72**	**1.46**	**2.03**	**1.68**	**2.36**	**2.27**

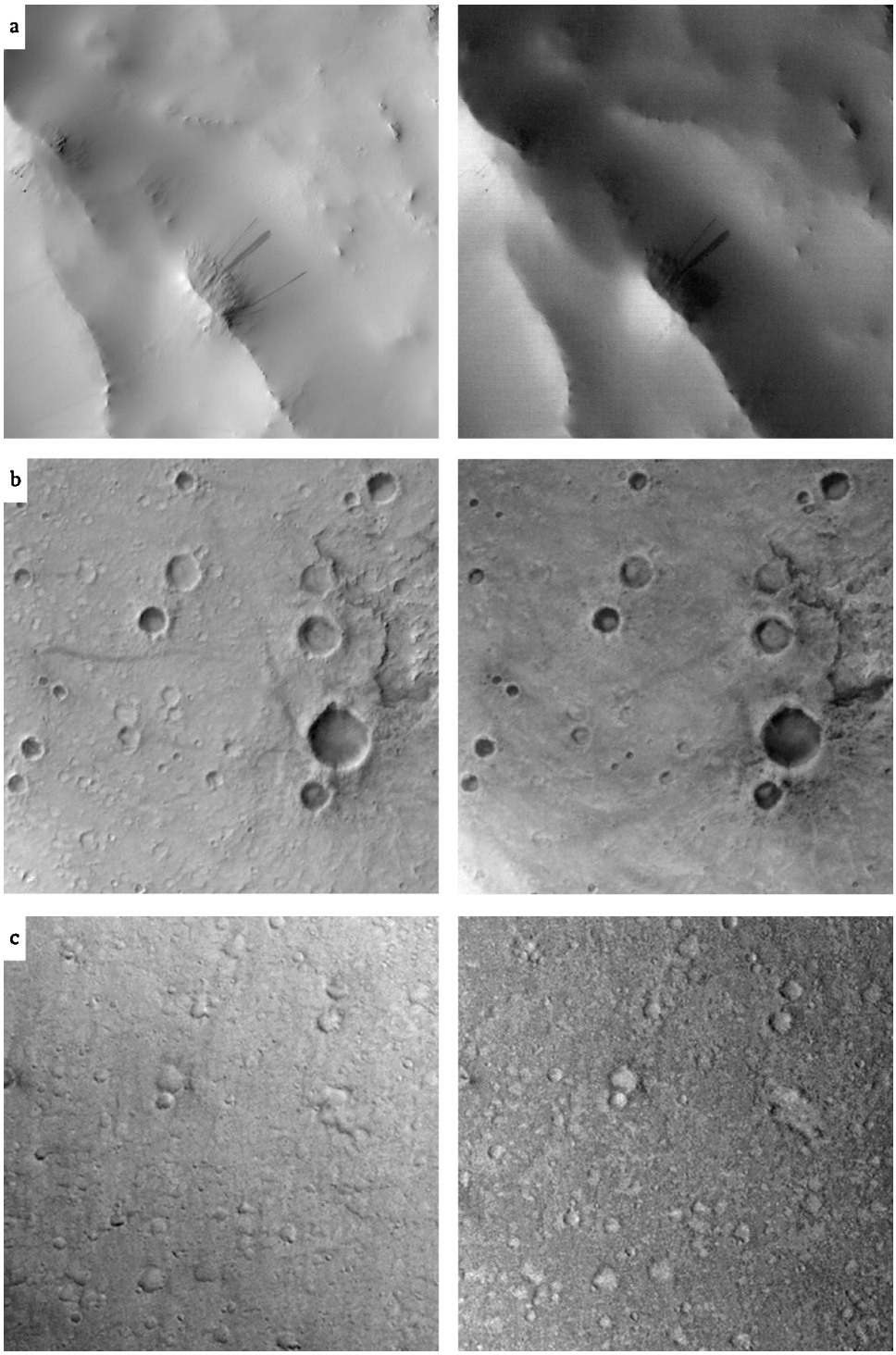

Fig. 8. Examples of a) 100% agreement between classifiers that a landscape features has changed; b) 50% agreement, and; c) 100% agreement that no landscape feature has changed.

participants use it by default. It might be used by too frequently in the case that participants will use it by default, so that it remains within the analysis set and prevents other images from being analysed.

Third and finally, implicit to the study is that participants marked features that had changed. This is important to acknowledge in any analysis; results do not reflect (expert and novice) participants' ability to identify any one individual feature because, in the context of the study presented, participants were not asked to mark features unless the feature had changed. Results instead demonstrate the relative success with which participants identified features that had changed, and differences would suggest how easy certain changes are to spot than others.

The results and their implications are now discussed accordingly, with respect to the study's aims and objectives.

4.1. Difference between features

Changes in some types of geological feature appear to be easier than others to spot; this is most clear in Fig. 9, which shows that dust devil tracks and slope streaks are easier to identify, since the consensus for changes in these features is highest. The emergence of differences between novice and expert classifiers only when asked to identify the landscape feature that has changed suggests a subtle affect of expertise on this task. Given the literature concerning visual inspection and

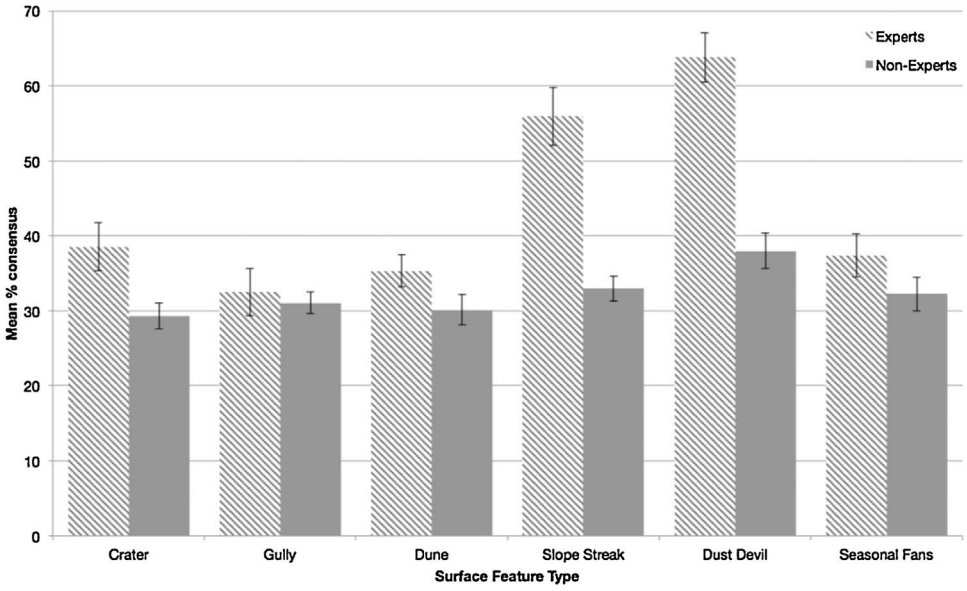

Fig. 9. Bar chart comparing the mean% consensus for different surface features between experts (Table 2) and novices (Table 3) with standard error bars to illustrate the significance of differences.

Table 4
Comparison of change detection performance.

Feature	% Images marked by experts	% Images marked by novices	% Agreement of novices with experts for images with change	% Disagreement of novices with experts for images with change
Impact crater	4.61	6.91	25	75
Gully	1.15	6.68	0	100
Dune	4.38	8.99	21.0	79.0
Slope streak	6.68	9.91	24.1	75.9
Dust Devil Track	11.52	9.91	52	48
Seasonal Fan	2.76	4.15	16.7	83.3

change detection, results are likely to have been confounded by factors beyond the scope of the paper and attributes of the individual image pairs; however, experts and novices were deliberately shown the same image set to mitigate these effects having an impact when classifications are compared.

4.2. Difference between experts and novices

There was an insignificant difference in the identification of change between novices and experts; differences only began to emerge when they were asked to classify features. The feature with the highest consensus amongst both experts and novices was dust devils, which might be due to their comparatively simple morphology. The low level of agreement within and between experts and novices is surprising; the highest level of agreement, amongst experts, is 64% for dust devils. This raises questions over the "expertise" required for this task, and the genuine difficulty of spotting changes in geological features from satellite imagery, and perhaps points to the need for a more detailed tutorial on the task than was provided for this study.

4.3. Relation to previous work

The results presented add nuance to the conclusions of Robbins et al. (2014) and suggest that there are subtle limitations in the tasks Citizen Scientists can perform to the standard of professional scientists. Whilst the strengths of Citizen Science are widely discussed, comparatively little is said about its limitations, which increase with the complexity of the tasks Citizen Scientists undertake. This paper contributes an appraisal of the Citizen Science approach for the more complex task of change detection and results especially suggest that the task of change detection places new demands upon untrained volunteers, which they do not appear to meet for many geological features.

Table 5
Results of test for normal distribution.

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig	Statistic	df	Sig
Experts	0.352	434	0.000	0.320	434	0.000
Novices	0.414	434	0.000	0.174	434	0.000

[a] Lilliefors Significance Correction.

4.4. Implications for the Earth Observation field

The levels of agreement between experts and novices presented in Table 4 cast doubt on the range of tasks researchers can reasonably expect novices to perform to a professional level. Change detection is ostensibly a task that can be carried out with minimal training, but the results presented suggest otherwise. We may therefore have to re-evaluate our expectations of the Citizen Science approach for contributing to change detection studies, and untrained volunteers' performance with a geological change detection task on remote sensing imagery.

4.5. Implications for Citizen Science and Volunteered Geographic Information

The results presented concern Citizen Science researchers because they raise questions regarding the number of volunteers required for meaningful results when annotating changes between images. The first

question the study set out to answer was whether volunteers can be given the task of looking for changes on the surface of Mars and produce results that are comparable to those produced by Planetary Scientists, on the premise that it is conceptually similar to 'Spot the difference'; however, the answer to this is clearly not as simple as might be assumed, as we found that differences emerged when classifiers were asked to classify feature type.

The debate over how to handle input data quality surfaced in the study presented. Whilst researchers using the Citizen Science approach tend to espouse the use of forced choice in task design, in the interest of the data collected being useful, the experience of this study points to tensions between image quality and scientists' trust and use of volunteered data analysis, and a trade off between the engagement of the volunteer and goals of the project. Whilst the project this study focused on collected data for the training of a change detection algorithm, the interaction of the discussion of this point will required with a second iteration of the algorithm.

4.6. Limitations and future work

Future studies should continue to investigate how and why the data for different feature types varies as this could have implications for the extension of the crowd sourcing approach from crater counting to other landscape features. The differences found in the consensus for different features types – both within and between experts and novices – suggest that the number of novice volunteers required to produce data comparable to a group of experts is inconsistent and the reasons for this remain unclear.

The set of images used in this study was also limited, and further work is required with more imagery, to examine the interplay between a change detection algorithm and novice classifications in more detail, and if and how they can work together effectively and efficiently. Work on the automatic change detection pipeline presented here is ongoing and will be published in more detail.

Further work can also investigate differences between novices' and experts' performance with a simplified task concerned only with feature identification. Such a study would go some way to determine whether the differences found in the present study can be attributed to the detection of change in a feature or whether they are also partly explained by the morphology of the features themselves.

5. Conclusion

This paper has explored the potential of novices to detect change in remotely sensed imagery of geomorphological features to a standard comparable with a group of expert classifiers, to further understand the strengths and limitations of the crowdsourcing approach within the fields of Earth Observation and Planetary Science. To do this, a Citizen Science project was created and tested on a group of Planetary Scientists before was made public. The task showed participants two images of the same location on the surface of Mars at two different point in time and where they marked change they were asked to select the feature type they had marked.

In short, the study found similarities in novice and expert identification of image pairs that have changed, but differences in the classification of the type geological features both within and between experts and novice classifiers, which should inform future work into the suitability of crowd-sourcing for similar scientific aims. It has demonstrated that consensus on changes in some geological features is much higher, suggesting that the analysis of these features is much more suited to crowd sourcing than others. Future work will investigate the underlying causes of these differences, to determine the effect of factors suggested by this study, such as the inherent complexity of feature morphology, prior exposure to imagery and whether a more detailed training in the task can improve classification consensus.

Acknowledgements

The research leading to these results received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under iMars grant agreement number 607379 and STFC MSSL Consolidated grant number ST/K000977/1. The authors also gratefully acknowledge the funding of Europlanets for the RPIF-3D workshop and its participants for their time and feedback on 'Mars in Motion', in addition to the volunteers who provided the data presented here. This publication uses data generated via the Zooniverse.org platform, development of is which funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

References

Bandeira, L., Marques, J.S., Saraiva, J., Pina, P., 2013. Advances in automated detection of sand dunes on Mars. Earth Surf. Processes Landf. 38 (3), 275–283.
Bell III, J.F., Malin, M.C., Caplinger, M.A., Fahle, J., Wolff, M.J., Cantor, B.A., James, P.B., Ghaemi, T., Posiolova, L.V., Ravine, M.A., Supulver, K.D., Calvin, W.M., Clancy, R.T., Edgett, K.S., Edwards, L.J., Haberle, R.M., Hale, A., Lee, S.W., Rice, M.S., Thomas, P.C., Williams, R.M.E., 2013. Calibration and performance of the mars reconnaissance orbiter context camera (CTX). Int. J. Mars Sci. Explor. 8 (April), 1–14.
Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J., 2009. Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59 (11), 977–984.
Bowyer, A., Lintott, C., Hines, G., Allan, C., Paget, E., 2015. Panoptes, a Project Building Tool for Citizen Science. Association for the Advancement of Artificial Intelligence (AAAI) Conference on Human Computation and Crowdsourcing (HCOMP 15). The AAAI Press, San Diego, CA, U.S.A, pp. 1–2.
Bue, B.D., Stepinski, T.F., 2007. Machine detection of martian impact craters from digital topography data. IEEE Trans. Geosci. Remote Sens. 45 (1), 265–274.
Castelhano, M.S., Heaven, C., 2010. The relative contribution of scene context and target features to visual search in scenes. Atten. Percept. Psychophys. 72 (5), 1283–1297.
Chang, T.-W., Kinshuk, Chen, N.-S., YU, P.-T., 2012. The effects of presentation method and information density on visual search ability and working memory load. Comput. Educ. 58 (2), 721–731.
Christensen, P.R., Jakosky, B.M., Kieffer, H.H., JRHYM, Mehall, G.L., Silverman, S.H., Ferry, S., Caplinger Ravine, M.M., 2004. Mars odyssey. In: Russell, C.T. (Ed.), The Thermal Emission Imaging System (Themis) for the Mars 2001 Odyssey Mission. Dordrecht: Springer, Netherlands, pp. 85–130.
Foody, G.M., See, L., Fritz, S., VAN DER VELDE, M., PERGER, C., SCHILL, C., BOYD, D.S., 2013. Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project. Trans. GIS 17 (6), 847–860.
Gault, D.E., 1970. Saturation and equilibrium conditions for impact cratering on the lunar surface: criteria and implications. Radio Sci. 5 (2), 273–291.
Gwinner, K., Jaumann, R., Hauber, E., Hoffmann, H., Heipke, C., Oberst, J., Neukum, G., Ansan, V., Bostelmann, J., Dumke, A., Elgner, S., Erkeling, G., Fueten, F., Hiesinger, H., Hoekzema, N.M., Kersten, E., Loizeau, D., Matz, K.D., Mcguire, P.C., Mertens, V., Michael, G., Pasewaldt, A., Pinet, P., Preusker, F., Reiss, D., Roatsch, T., Schmidt, R., Scholten, F., Spiegel, M., Stesky, R., Tirsch, D., Van Gasselt, S., Walter, S., Wählisch, M., Willner, K., 2016. The High Resolution Stereo Camera (HRSC) of Mars Express and its approach to science analysis and mapping for Mars and its satellites. Planet. Space. Sci. 126 (July), 93–138.
Haklay, M., 2010. How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Enviro. Plan. B: Plan. Des. 37 (4), 682–703.
Jaumann, R., Neukum, G., Behnke, T., Duxbury, T.C., Eichentopf, K., Flohrer, J., Gasselt, S.V., Giese, B., Gwinner, K., Hauber, E., Hoffmann, H., Hoffmeister, A., Köhler, U., Matz, K.D., Mccord, T.B., Mertens, V., Oberst, J., Pischel, R., Reiss, D., Ress, E., Roatsch, T., Saiger, P., Scholten, F., Schwarz, G., Stephan, K., Wählisch, M., 2007. The high-resolution stereo camera (HRSC) experiment on Mars Express: instrument aspects and experiment conduct from interplanetary cruise through the nominal mission. Planet. Space Sci. 55 (7–8), 928–952.
Kim, J.R., Muller, J.-P., Van Gasselt, S., Morley, J.G., Neukum, G., 2005. Automated crater detection, a new tool for mars cartography and chronology. Photogramm. Eng. Remote Sens. 71 (10), 1205–1217.
Kirchoff, M., Sherman, K., Chapman, C., 2011. Examining lunar impactor population evolution: additional results from crater distributions on diverse terrains. EPSC-DPS Joint Meeting 2011 1587.
Li, B., Ling, Z., Zhang, J., Wu, Z., 2015. Automatic detection and boundary extraction of lunar craters based on LOLA DEM data. Earth Moon Planets 115 (1), 59–69.
Lloyd, R., Hodgson, M.E., 2002. Visual search for land use objects in aerial photographs. Cartogr. Geogr. Inf. Sci. 29 (1), 3–15.
Malin, M.C., Edgett, K.S., Cantor, B.A., Caplinger, M.A., Danielson, G.E., Jensen, E.H., Ravine, M.A., Sandoval, J.L., Supulver, K.D., 2010. An overview of the 1985–2006 mars orbiter camera science investigation. Int. J. Mars Sci. Explor. 5 (Jan), 1–60.
Rensink, R.A., 2002. Change detection. Annu. Rev. Psychol. 53 (1), 245–277.
Robbins, S.J., Antonenko, I., Kirchoff, M.R., Chapman, C.R., Fassett, C.I., Herrick, R.R., Singer, K., Zanetti, M., Lehan, C., Huang, D., Gay, P.L., 2014. The variability of crater

identification among expert and community crater analysts. Icarus 234 (15 (May)), 109–131.
Salk, C.F., Sturn, T., See, L., Fritz, S., 2016. Limitations of majority agreement in crowdsourced image interpretation. Trans. GIS (n/a-n/a).
See, L., Fritz, S., Dias, E., Hendriks, E., Mijling, B., Snik, F., Stammes, P., Vescovi, F.D., Zeug, G., Mathieu, P.P., Desnos, Y.L., Rast, M., 2016. Supporting Earth-Observation Calibration and Validation: a new generation of tools for crowdsourcing and citizen science. IEEE Geosci. Remote Sens. Mag. 4 (3), 38–50.
See, J., 2012. Visual Inspection: A Review of the Literature: SAND2012-8590. Sandia National Laboratories(Available online from: http://prod.sandia.gov/techlib/access-control.cgi/2012/128590.pdf [Accessed: 22/03/2016]).
Sidiropoulos, P., Muller, J.P., 2015. On the status of orbital high-resolution repeat imaging of Mars for the observation of dynamic surface processes. Planet. Space Sci. 117, 207–222.
Sidiropoulos, P., Muller, J.P., 2016. Big data from Mars: design of global planetary data analysis tools. In: 2016 Conference on Big Data from Space − BiDS'16. Santa Cruz de Tenerife, Spain. pp. 320–323.
Stepinski, T.F., Collier, M.L., 2004. Extraction of Martian valley networks from digital topography. J. Geophys. Res.: Planets 109 (E11) (n/a-n/a).
Tar, P.D., Thacker, N.A., 2016. Automated Feature Quanitifcation for Planetary Surfaces Verses Human Subjectivity: Tina Memo No. 2016-010. University of Manchester (Available online from: http://www.tina-vision.net/docs/memos/2016-008.pdf [Accessed: 02/09/2016]).
Wagstaff, K.L., Panetta, J., Ansar, A., Greeley, R., Hoffer, M.P., Bunte, M., Schrghofer, N., 2012. Dynamic landmarking for surface feature identification and change detection. ACM Trans. Intell. Syst. Technol. 3 (3), 1–22.
Wagstaff, K.L., Doran, G.B., Kiran, R., Mandrake, L., Schorghofer, N., Stanboli, A., 2015. Landmark classification and content-Based search for mars orbital imagery. Second Planetary Data Workshop. (pp).
Warner, N.H., Gupta, S., Calef, F., Grindrod, P., Boll, N., Goddard, K., 2015. Minimum effective area for high resolution crater counting of martian terrains. Icarus 245 (1 January), 198–240.
Ye, G., Liu, D., Jhuo, I.H., Chang, S.F., 2012. Robust late fusion with rank minimization. 2012 IEEE Conference on Computer Vision and Pattern Recognition 3021–3028.
Zhao, Y.Y., Gong, P., Yu, L., Hu, L.Y., Li, X., Li, C.C., Zhang, H., Zheng, Y., Wang, J., Zhao, Y.C., Cheng, Q., Liu, C., Liu, S., Wang, X.Y., 2014. Towards a common validation sample set for global land-cover mapping. Int. J. Remote Sens. 35 (13), 4795–4814.